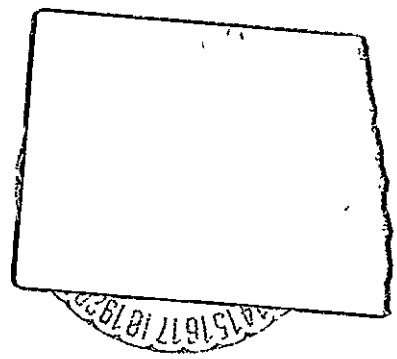
LARS Information Note
070676

Determining Land Use Patterns Through Man-Machine Analysis of LANDSAT Data ... A Tutorial Simulation

by Stevan J. Kristof, James D. Russell, Tina K. Cary, Bruce M. Lube, and Richard A. Weismiller



The Laboratory for Applications of Remote Sensing
Purdue University West Lafayette, Indiana
1976

LARS Information Note 070676

DETERMINING LAND USE PATTERNS THROUGH MAN-MACHINE ANALYSIS OF LANDSAT DATA -- A TUTORIAL SIMULATION

by
Stevan J. Kristof
James D. Russell
Tina K. Cary
Bruce M. Lube
Richard A. Weismiller

The Laboratory for Applications of Remote Sensing
Purdue University
West Lafayette, Indiana
47906

This work was supported by the National Aeronautics and Space Administration under contract NAS9-14016 and NAS9-14970.

T-1039/4

DETERMINING LAND USE PATTERNS THROUGH MAN-MACHINE ANALYSIS OF LANDSAT DATA -- A TUTORIAL SIMULATION

by Stevan J. Kristof, James D. Russell, Tina K. Cary, Bruce M. Lube, and Richard A. Weismiller

This publication is designed as a simulation to show you typical steps in the analysis of remotely-sensed data for determining land use patterns. The example uses numerically-oriented pattern recognition techniques and emphasizes the respective roles of the analyst (man) and the computer (machine) in the analysis process.

PREREQUISITES

The material is intended for persons who have experience in land-use management and a general background in remote sensing. The remote sensing background can be gained by means of the following educational materials or their equivalent:

LARSYS Educational Package[1]

- Unit I An Introduction to Quantitative Remote Sensing
- Unit II LARSYS Software System: An Overview

Fundamentals of Remote Sensing Minicourse Series[2]

- Remote Sensing: What Is It?
- The Physical Basis of Remote Sensing
- Spectral Reflectance Characteristics of Vegetation
- Spectral Reflectance Characteristics of Earth Surface Features

The principles and techniques described in this simulation apply to numerical analysis procedures in general. LARSYS is used as an example of a numerical analysis software system for this simulation.

---

[1]The LARSYS Educational Package may be obtained from the System Services Manager, Laboratory for Applications of Remote Sensing, 1220 Potter Drive, West Lafayette, Indiana, 47906.

[2]The Minicourse Series may be obtained from G. W. O'Brien, Continuing Education Administration, 116 Stewart Center, Purdue University, West Lafayette, Indiana, 47907.

PREFACE

The purpose of this simulation is NOT to train you to be a remote sensing data analyst, but instead to give you an overview and understanding of how data are analyzed to determine land use. Our purpose here is analogous to teaching you how your automobile operates, but not teaching you how to drive it.

It should be pointed out that the experience of the analyst is a very important factor in the man-machine interaction described in this simulation. Stevan Kristof who generated this analysis has a soil science background and over ten years of experience with computer-aided analysis of multispectral data.

GENERAL OBJECTIVE

Upon completion of this simulation, you should be able to describe the process of analyzing remotely-sensed data using numerical analysis techniques. Your description should include the nature of the interaction among input (data), man (analyst) and machine (computer), and the product (results) of each step in the process.

ACKNOWLEDGEMENTS

The authors wish to thank Philip H. Swain, Program Leader for Data Processing and Data Research at LARS, and John C. Lindenlaub, Program Leader for Technology Transfer at LARS, who served as subject-matter advisors and provided valuable input for this simulation. Numerous other who reviewed rough drafts of this material made important comments and suggestions which were incorporated into the simulation. A special thanks goes to John Berkebile whose Forestry Application Simulation of Man-Machine Techniques for Analyzing Remotely Sensed Data served as a model for this publication.

TABLE OF CONTENTS

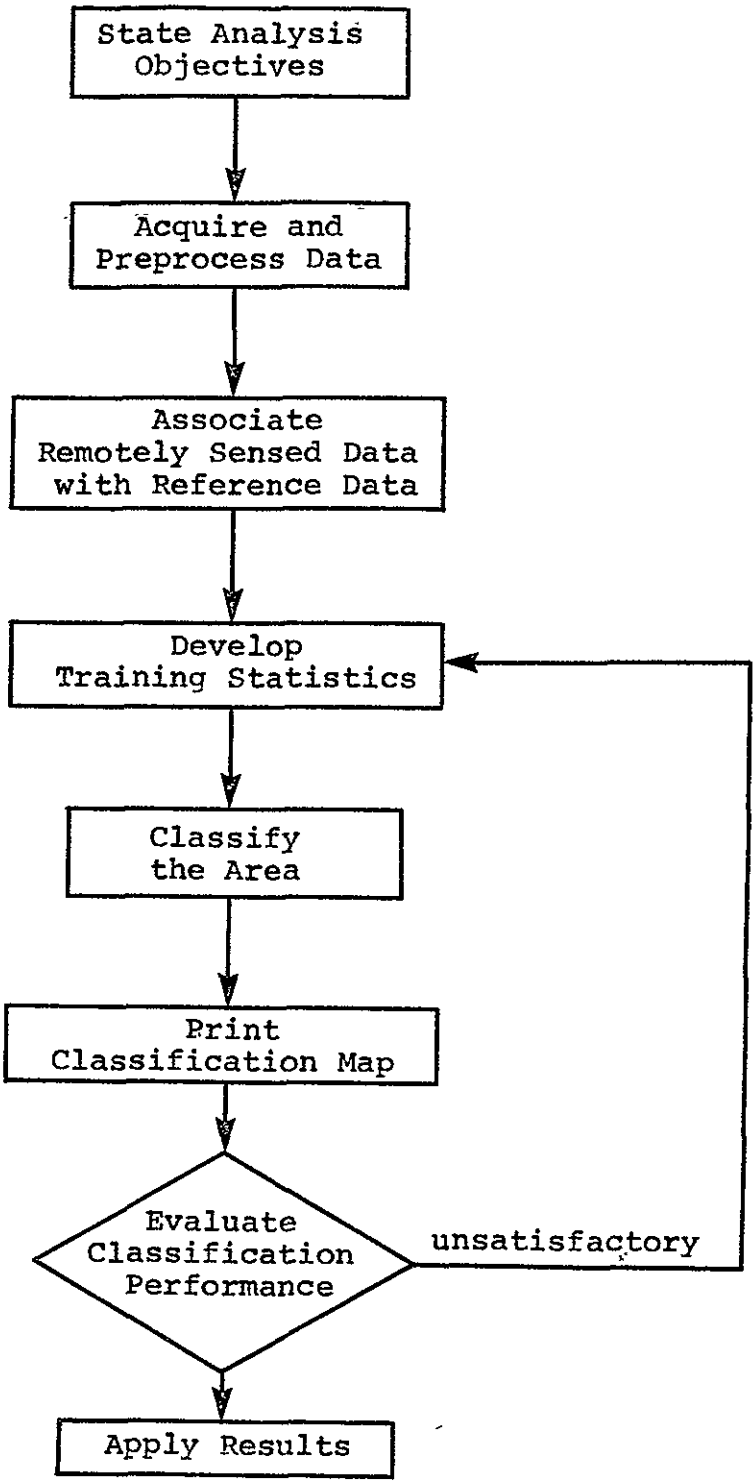


Figure 1 Typical Numerical Analysis Flowchart for Land Use Determination

OVERVIEW

Inventorying and monitoring land resources are fundamental to the interests of national, state and local governments. In response to Federal legislation designed to give impetus to programs in land resource management at the state level, NASA has as one of its objectives:

> "To define, design and develop, and demonstrate a Regional Land Resources Inventory System(s), which utilizes, in part, information extracted from remotely sensed data for the inventory and monitoring of regional land resources in support of resource development and use management."[1]

The numerical analysis of remotely sensed data is a dynamic process which requires an interaction between man (analyst) and machine (computer). The process relies on a documented technology of remote sensing analysis and upon judgements and insights by the analyst. Remote sensing techniques allow you to survey large areas with a minimum of time and cost. The computer can be "trained" to produce general land use maps. A typical analysis sequence is shown in Figure 1. Even though it is shown here as basically a linear process, all of the steps are interrelated. At any step in the analysis, interpretation of the results of that step can lead the analyst to conclude that he should go back to a previous step and revise his procedure.

Machine-aided processing of remotely sensed data allows those interested in surveying relatively large areas to do so more quickly and with less cost than other methods. Once familiarity with an area and with the spectral responses of cover types is gained, a monitoring procedure is made more feasible. Tarnocai and Kristof[2] used remote sensing techniques to survey the MacKenzie River Delta area in Canada. LANDSAT I multispectral scanner data was correlated with aerial photographs and ground observations to classify this large, relatively inaccessible area. In a similar effort, the General Land Office in Austin, Texas, is currently engaged in inventorying and monitoring resources of the Texas coastal zone.[3] Other investigators are also using remote sensing when a rapid, relatively

---

[1] Regional Land Resources Inventory and Monitoring Applications System Verification Test Project, Preliminary Plan, NASA/JSC, Houston, Texas, 1973

[2] Tarnocai, C. and S. J. Kristof. "Computer-Aided Classification of Land and Water Bodies Using ERTS Data - MacKenzie Delta Area, N. W. T." LARS Information Note 031875, 1975. Arctic and Alpine Research, Vol. 8, No. 2, 1976.

[3] Armstrong, Robert (principal investigator), "Development and Application of Operational Techniques to Inventory and Monitor Resources and Uses in Texas Coastal Zone," December, 1973.

inexpensive method of monitoring change in a given area is needed. Since the LANDSAT satellites repeat their cycle every eighteen days, current data is continually available. The progress of flood water, changes in the integrity of a coastal area, and large area changes in the use of land resources are just a few ways remotely sensed multispectral scanner data can and are being used. The project from which this tutorial simulation was derived examined portions of the Texas Coastal Zone. However, for simplicity, the analysis presented here is from an area covered by just one U.S.G.S. 7½ minute quadrangle map (Pass Cavallo).

When an analysis problem in remote sensing is conceived, certain steps are followed by the analyst. The first step is to state the analysis objectives. To do this, the analyst must determine the geographic area of interest, the general cover types present and the nature of the application to which the results will be applied. An additional component which is often included in the analysis objective is the desired classification accuracy for initial estimates of land use. An example would be to "determine the percentage of the Pass Cavallo quadrangle (Texas coastal zone) represented by each of these ground cover types: sand, swales, shrubs, bare areas, shallow water and deep water and other, with 85% accuracy."

Remotely sensed data can be acquired via aircraft or satellite. The platform used and type of data collected will be determined in part by the analysis objectives. The usefulness of any data may be affected by atmospheric conditions (haze, cloud cover) and system problems such as striping. Prior to analysis, data is usually preprocessed. Preprocessing is carried out by the computer and usually includes rotating the image so that the vertical columns are aligned in a north-south orientation and changing the scale to correspond to maps available to the analyst.

Next, the remotely-sensed data is correlated with the available reference data. The reference data might include USGS topographic maps, aerial photographs, and actual ground observations. Since each LANDSAT satellite covers the entire earth every eighteen days, the analyst can generally choose the time of the year most suitable for mapping the cover types of interest.

Training areas within the data are then selected and statistics calculated for these areas. The training areas are selected to contain typical examples of each cover type of interest and the coordinates of these areas are supplied to the computer in order to "train" it to classify unknown data points. There are some general selection criteria to aid the analyst in choosing training areas, but successful training area selection relies heavily on the analyst's previous experience and knowledge of the location based upon direct observation and reference data. When training

areas have been selected the analyst may call upon the computer to use information from more than one channel, or wavelength band, to "cluster" the data in the training areas into spectrally distinct "cluster classes." Field Description Cards for each cluster class are used as input by the computer to calculate the mean vector and covariance matrix for each class. These numbers define the "training" statistics.

Next, the area of interest is "classified." That is, every data point in the area is assigned to one of the "training" classes. The rule used for assigning points to classes is to assign the data point to the class to which it has greatest probability of belonging.

In the next step, a map is printed by the computer. The computer is instructed to indicate (by printing a blank) any data points that do not have a certain probability of belonging to the class to which they are assigned. This procedure results in a map that shows regions of the data unlike any of the training classes. From these regions, additional training data are selected. The results are evaluated by comparison with available reference data.

As indicated earlier, numerical analysis of multispectral scanner data is a dynamic process with each step providing feedback to the previous step. For simplicity, the process is shown here as a linear sequence. In reality, the analyst has all steps in mind before he actually begins an analysis. He may also refer back to previous steps and modify his procedure as the analysis continues.

Now that we have looked at an overview of the entire process, let's go back and look at each step in more detail. You will want to refer frequently to Figure 1 to keep in mind exactly where you are during the discussion of the numerical analysis process.

Figure 2 The analyst and user working together to develop analysis objectives

SECTION I STATE ANALYSIS OBJECTIVES

Upon completion of this section, you should be able to:

1. Name the four components of an analysis objective.
2. Write an analysis objective incorporating these four components.

The first and one of the most important steps in the numerical analysis process is stating the analysis objectives. What is the problem to be solved? What information do you need to solve the problem? What are you going to do with the results of the analysis?

For example, the purpose of the analysis could be to assist in choosing recreational sites in an area along the coast of Texas. Information on present land use is needed to make such a decision. In this case, the analysis objective might be:

> "Determine the land use in the Texas Coastal Zone with 80% accuracy in order to select the most desirable site for the location of new recreational areas."

If the Department of Highways is planning to build a new highway in the area, the analyst might want to:

> "Generate a soil type map of the Point Comfort quadrangle to aid in the selection of the most desirable right-of-way for a new highway."

In order to provide the proper environment for wildlife, the analysis might have as its objective to:

> "Locate suitable cover types and wetlands (habitat diversity) for wildlife species in the Austwell quadrangle as an aid in managing wildlife habitats."

The essential components of an analysis objective are:

Location What portion of the earth's surface is of interest? It may be a relatively small area (several hundred hectares using airborne multispectral scanners) or a relatively large

area (thousands or millions of hectares using multispectral scanner data from satellite-borne systems).

Cover Types What types of ground cover are of interest to you? Are you interested in woodlands, agriculture, rangeland, pasture, barren land and marshes? Are you interested in water such as shallow water, deep water, fresh water or salt water?

Applications How will the analysis output be used? To map drainage patterns? To locate potential port or harbor locations? To inventory a specific area and prepare a cover type map?

Classification Performance How accurate must the classification be in order to be of help to you? Would a classification performance of 65% be acceptable or do you need to have approximately 90% accuracy? The level of accuracy that can be obtained depends upon many factors: the level of detail desired, time of the year data were collected, the analyst's training and skill, particular region being mapped, and other variables. An indication of the accuracy required by the user is very useful to the analyst as it provides a point of reference to which he can compare and assess his analysis.

The analysis objective for this simulation was determined by the analyst working with the person who would eventually use the information derived from the analysis. In some cases, the objectives are given to the analyst by the user or the agency needing the information. The analysis objective was stated as follows:

> "Produce a detailed classification map of the Pass Cavallo quadrangle (see Figure 3) using computer-assisted analysis of LANDSAT-1 data. The cover types to be mapped are: various water classes, range land, marshes, brushland, barren land, and man-made features. Areas of the above cover types greater than 2.5 hectares in extent must be accurately identified. The results will be used to prepare a land use management scheme compatible with range land usage and wildlife habitat preservation."

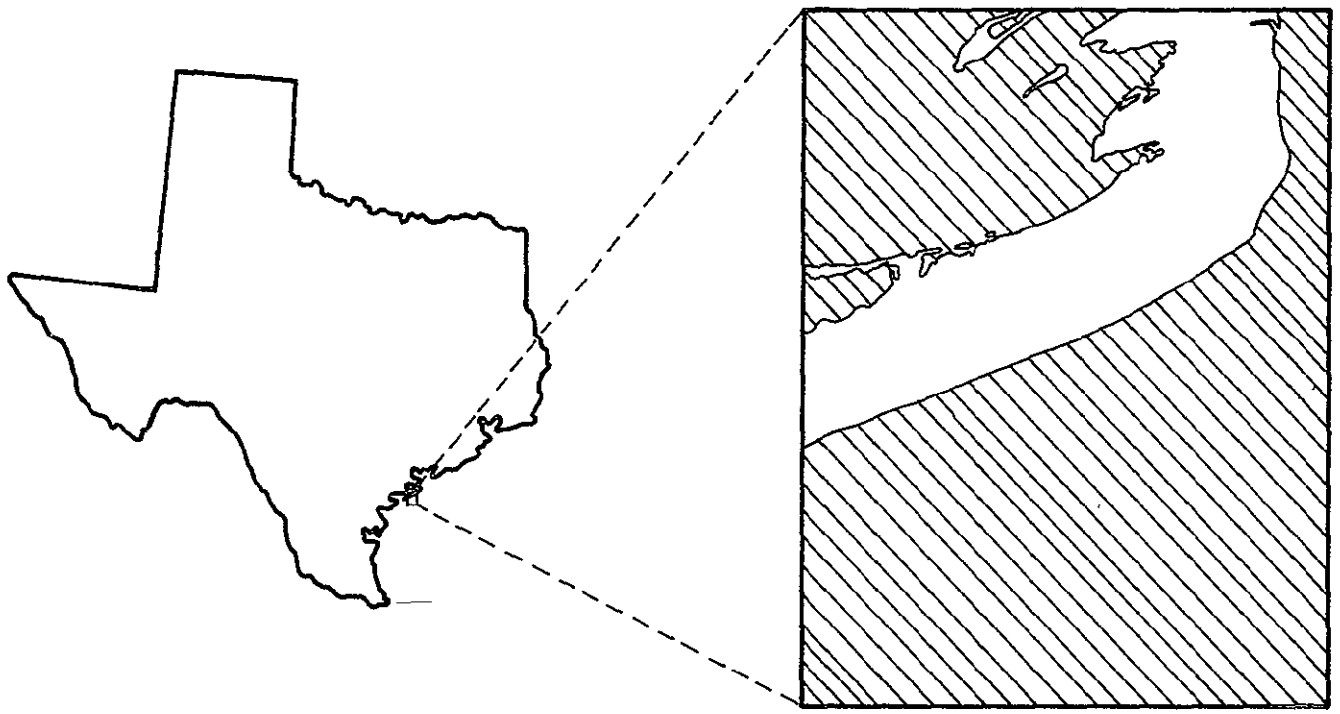


Figure 3 The Pass Cavallo quadrangle is a portion of the Texas Coastal Zone just south of Houston

Self-Check[1]

I-A. Name the four components of an analysis objective.

I-B. Write an analysis objective that you might use in solving a land use problem.

[1]Answers to all Self-Checks are given in Appendix B, page 51.

SECTION II ACQUIRE AND PREPROCESS DATA

Upon completion of this section, you should be able to:

1. Identify two systems for collecting multispectral scanner data for land use analysis.

2. Describe three data idiosyncrasies which might hinder analysis.

3. Describe two types of geometric corrections that might aid the analysis of LANDSAT data.

The analysis objective specifies the area to be studied. Multispectral data of the area can be obtained from sensors carried by aircraft or satellites. Typically, the total field-of view increases as the altitude of the data collection system increases. On the other hand, image resolution typically decreases as altitude increases, so there is less detail available from higher altitudes. The price the analyst pays for covering a larger area is usually poorer spatial resolution. If aircraft data is desired, commercial companies can be contracted to provide such data. Aircraft data, though generally more expensive for the user to acquire than satellite data, provides the advantage of allowing the analyst a wider selection of data formats (type and scale) and of specifying a time when weather conditions will not interfere with data collection over the area. Satellite data provides repetitive coverage over a larger area. A data distribution center for remotely sensed data is the EROS Data Center in Sioux Falls, South Dakota. LANDSAT data may be ordered from EROS in the following formats:

Image products such as
- 70mm negatives and positives
- 9 x 9 inch negatives, positives and color composites
- various size prints in black-and-white and color

Computer tapes containing the data in numerical format.

The type of data desired and the time of year for data collection are determined on the basis of the stated analysis objectives. For example, if you are examining the acreages planted in various crops, you want data collected at the time of growing season when the greatest spectral distinction among them occurs.

In carrying out an analysis, you must also be concerned about the data quality. In general, a greater level of analysis accuracy is possible when the original data is of the highest quality. A preliminary evaluation of digital data can be made by inspecting imagery created from the data. If after examining the imagery one finds the area of interest totally obscured by clouds, meaningful analysis of that data set will not be possible. Gross data characteristics which can significantly decrease the usefulness of a data set include haze, clouds and snow cover. Data sets can be screened for these characteristics by examining digital display images or grayscale printouts. (See Figure 4)

Figure 4 An example of clouds and their shadows

Sometimes "striping" will occur in the image. In the scanner system for LANDSAT I and II, six data lines are simultaneously recorded in each wavelength band each time the scanner mirror oscillates. A separate detector is used for each channel of each of these scan lines. If any of these detectors and their electronics are not properly matched or calibrated, the striping effect is noticeable in the imagery of that channel. A dramatic example is shown in Figure 5.

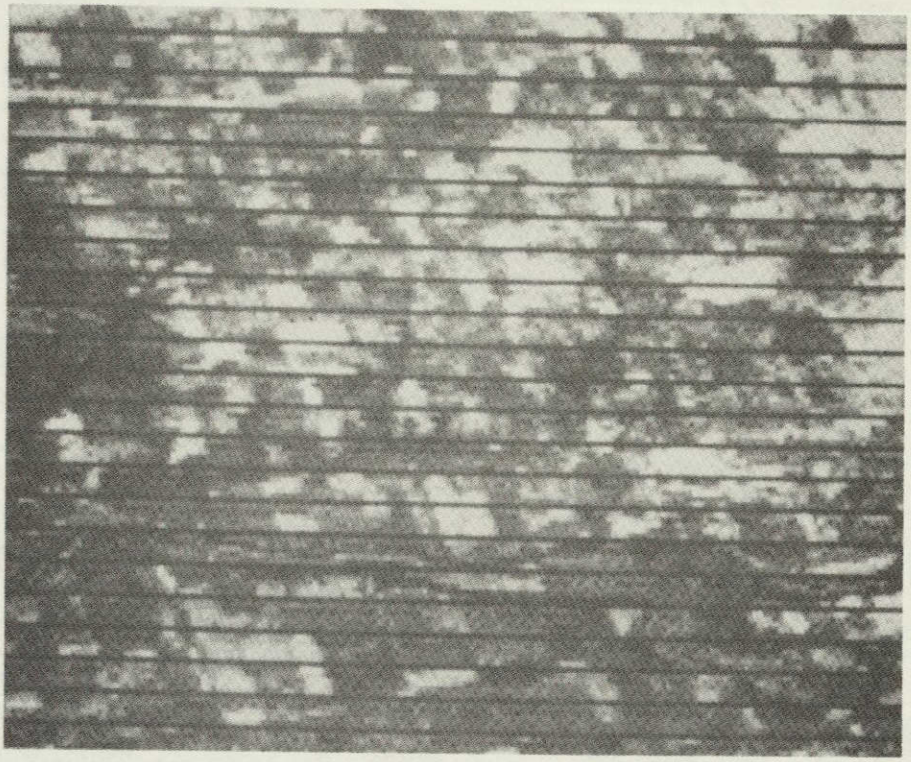
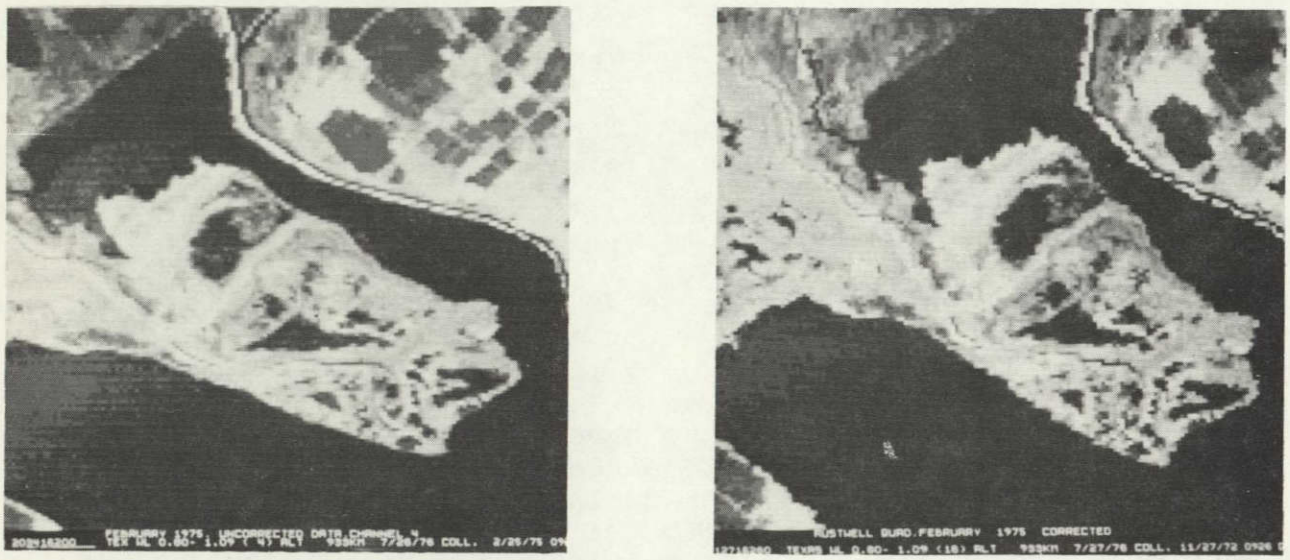


Figure 5 Example of striping of data

Even though striping appears to have seriously degraded the data, analysis may still be possible since striping usually occurs in only some of the channels. Since LANDSAT I and II have four channels, it may be possible to get meaningful analysis results when only two or three of the four channels are available.

Once a data set has passed an initial screening, the analyst may request some preprocessing of the data to make it easier for him to correlate the remotely sensed data with reference data. Preprocessing can facilitate the analyst's ability to interact with the data and produce output at the desired scale. One type of preprocessing (geometric correction) involves rotating and deskewing the image so that the vertical columns are aligned in a north-south orientation. If you imagine yourself trying to locate corresponding points between a LANDSAT image and an aerial photograph or map, you can see that this task would be much easier if the two images are oriented in the same direction. In another geometric correction, the data is rescaled. The scale of the LANDSAT data ouptut can be changed to allow the LANDSAT data to be overlayed on the reference data. For instance, in the analysis disucssed in this simulation, the analyst had 7½' USGS (quadrangle) maps available to him. It was therefore convenient to have the data preprocessed so that when printed on a standard computer line printer, the "computer map" was at the same scale as the USGS map, i.e. 1:24,000.

A more detailed look at the steps and mathematical basis of geometric correction can be found in LARS Information Note 103073, "Geometric Correction of ERTS-1 Digital Multispectral Scanner Data" by Paul E. Anuta. Pictorial illustrations of the effects of the rotation and rescaling operations are shown in Figure 6.



Uncorrected Corrected

Figure 6 Comparison of original and geometrically corrected and rotated LANDSAT imagery

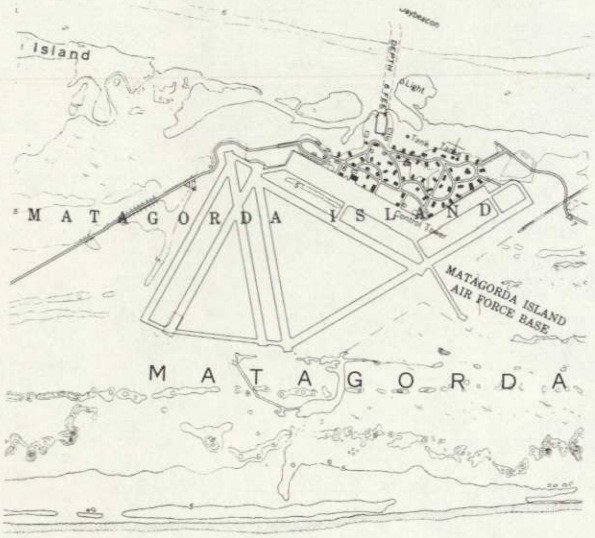
Having stated the analysis objectives, the analyst requests data, performs an initial data quality screening and, assuming the data is judged to be satisfactory, requests any preprocessing operations he feels will aid him in the analysis. He is then ready to begin associating the remotely sensed data with available reference data as described in the next section.

Self-Check

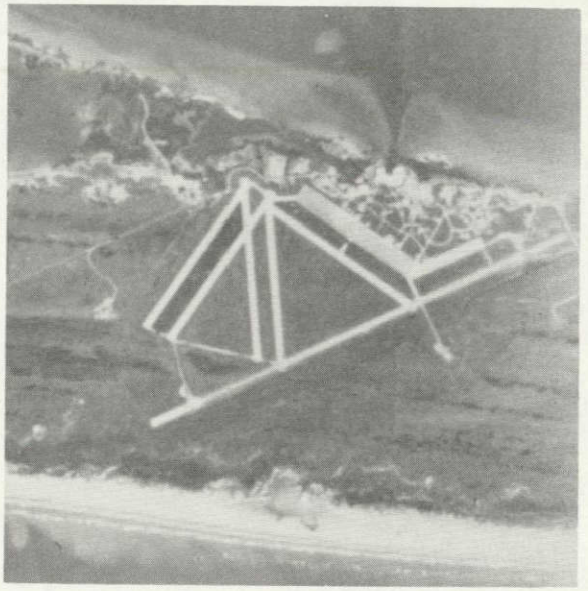
II-A. What are two systems for collecting multispectral scanner data for land use analysis?

II-B. Describe three data idiosyncrasies which might hinder your analysis.

II-C. What two types of geometric correction useful in the analysis of LANDSAT data were discussed in this simulation?



Topographic Map



Aerial Photograph

Field Observations

Figure 7 Typical examples of reference data

SECTION III ASSOCIATE REMOTELY-SENSED DATA WITH REFERENCE DATA

Upon completion of this section, you should be able to:

1. Describe three types of reference data.
2. Discuss the rationale for comparing remotely-sensed data with reference data.

In this step of the analysis, the analyst correlates the multispectral scanner data with available reference data. This not only allows him to gain familiarity with the geographic region being analyzed, but it will aid him in developing good training statistics and in evaluating the classification results later in the analysis process.

Reference data is any information which aids the analyst in his task. There are several types of reference data which may be available to assist the analyst in his work. Typical examples are aerial photographs, maps, previous analysis results and on-site observations.

One aid that is very useful is aerial photography. Four types of film are commonly used: black-and-white, black-and white infrared, color and color infrared. The type of film selected is determined by the nature of the application and the conditions under which it will be used. Infrared camera systems are often preferred for high altitude photography because of their haze penetration quality. Infrared films are also useful for enhancing differences between vegetation types.

Maps are often used as reference data. In this analysis a USGS topographic map proved to be a valuable piece of reference data. In the previous step, the data was preprocessed so that a computer line printer "map" had the same scale as the "topo" map. Thus, a simple overlay technique could be used to correlate the LANDSAT data with the map. By overlaying the computer printout of the area of interest on top of the quadrangle map, features of interest to the analyst can be located in the LANDSAT data and their row and column numbers noted.

A third type of reference data involves direct ground observation by someone trained to observe relevant characteristics of ground features of interest to the analyst. In certain geographic areas, ground observations have been recorded on maps. For the Texas Coastal Zone, there are maps known as B.E.G. (Bureau of Economic Geology) maps which are color-coded and vary in content. The content of the maps ranges from those

showing topographic and bathymetric features, to environmental and biological assemblages. One type illustrates land use patterns.

In this analysis, maps, aerial photographs and field observations were all used. USGS quadrangle maps and B.E.G. were very helpful to the analyst in getting an overview of the area. The aerial photography was used in an attempt to identify some types of ground cover since the aerial photos were taken at about the same times as the MSS data was gathered. Finally, the analyst visited the area and made ground observations in the Pass Cavallo quadrangle.

This section introduced the concept of reference data and described three types of reference data (aerial photographs, maps and ground observations) which are commonly used to aid the analyst in analyzing LANDSAT or other multispectral scanner data. The analyst correlated LANDSAT data of the Pass Cavallo area with the reference data available. Associating remotely-sensed data with reference data is largely an analyst operation. The analyst may be assisted by computer-generated images which facilitate his ability to interact with the data under study. The task of the analyst is to compare the patterns in the reference data with those in the images of the scanner data and select areas to be used for training the computer to recognize the ground features specified in the analysis objectives. These areas must contain the cover types of interest and should represent the variability of those cover types found in the ground scene. These are called "training areas" because once their cover types are identified, scanner data from those areas will be used to "train" the computer to recognize and identify these features throughout the area being studied.

Self-Check

III-A. What are three types of reference data?

III-B. Why does the analyst compare remotely-sensed data with reference data?

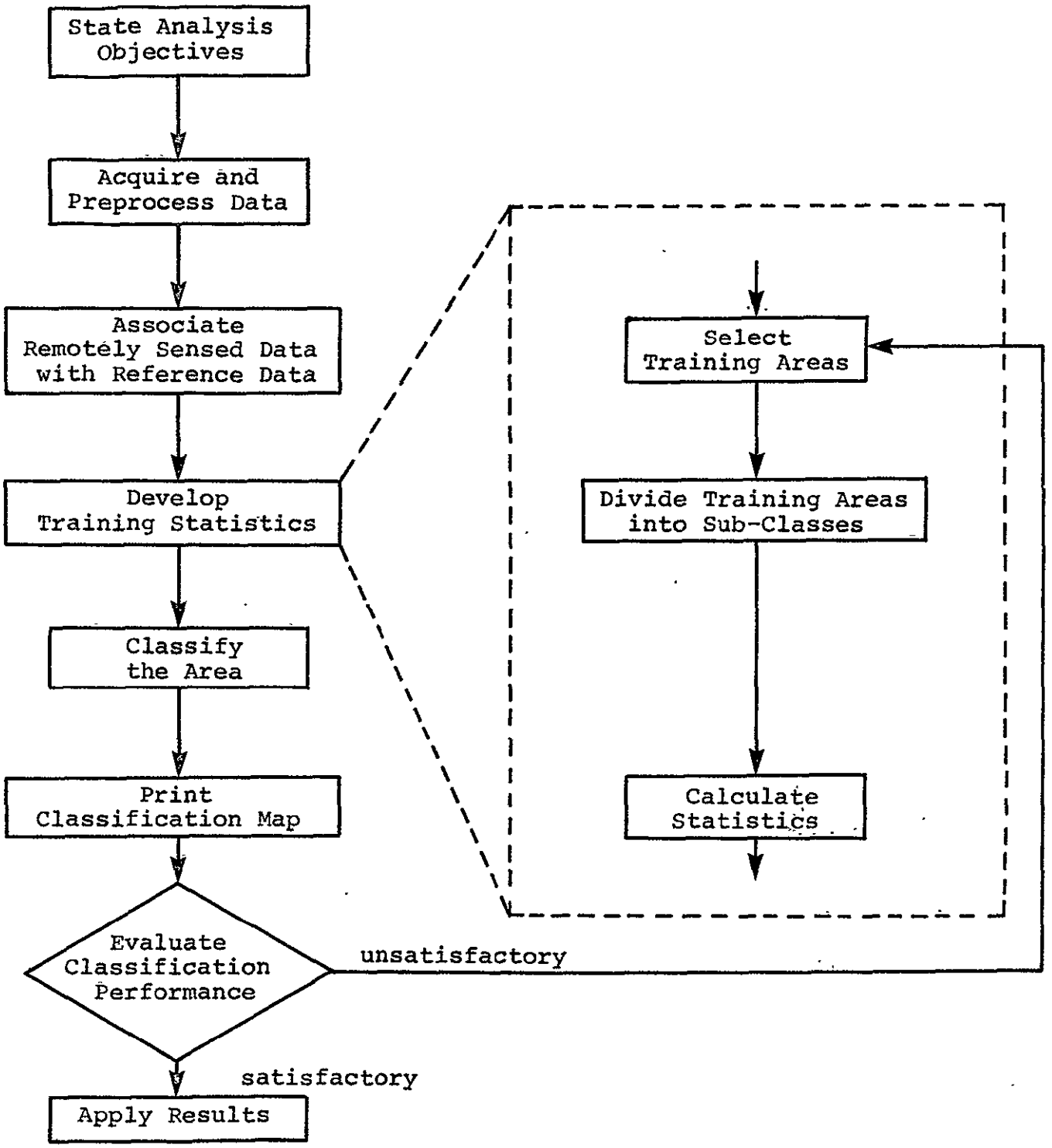


Figure 8 Typical Numerical Analysis Flowchart for Land Use Determination

SECTION IV DEVELOP TRAINING STATISTICS

Upon completion of this section, you should be able to:

1. Describe the procedure for selecting training areas. Include in your description factors of location, number and size.

2. Briefly describe clustering and how the analyst uses clustering to aid in the development of training statistics.

3. Discuss the nature and purpose of the statistics used in numerical analysis.

The process of developing the training statistics can be broken down into a series of three operations as shown in Figure 8. We will explain each of these in detail. Data collected by multispectral scanners and the reference data provide information which can be used to develop training statistics. The overall purpose of the training statistics is to "train" the computer to accurately classify the entire area of interest using pattern recognition techniques. Pattern recognition is a useful tool for identifying and classifying the ground scene. First, known ground cover types (patterns) are identified with the help of reference data. These patterns of known classification constitute the training patterns against which the patterns of unknown ground cover types are compared. A classification algorithm is used to "sort" the ground scene into various cover types. A data point of unknown type is classified according to a decision rule. The CLASSIFYPOINTS processor (see Appendix A, page 45) is based on the maximum likelihood decision rule, that is, a data point is assigned to the most probable or most likely ground cover type. For more detail see Pattern Recognition: A Basis for Remote Sensing Data Analysis by Philip H. Swain (LARS Information Note 111572).

Developing the training statistics for the area under investigation is the most critical and time consuming step for the analyst. More so than in any other part of the analysis, it requires interaction of man and machine. The classification performance is highly dependent upon how well the training data represents the entire area and how well the ground cover types can be distinguished.

Step 4-A Select Training Areas

As a first step toward the development of training statistics, the analyst usually determines the location, number and size of the training areas he is going to use. Several questions

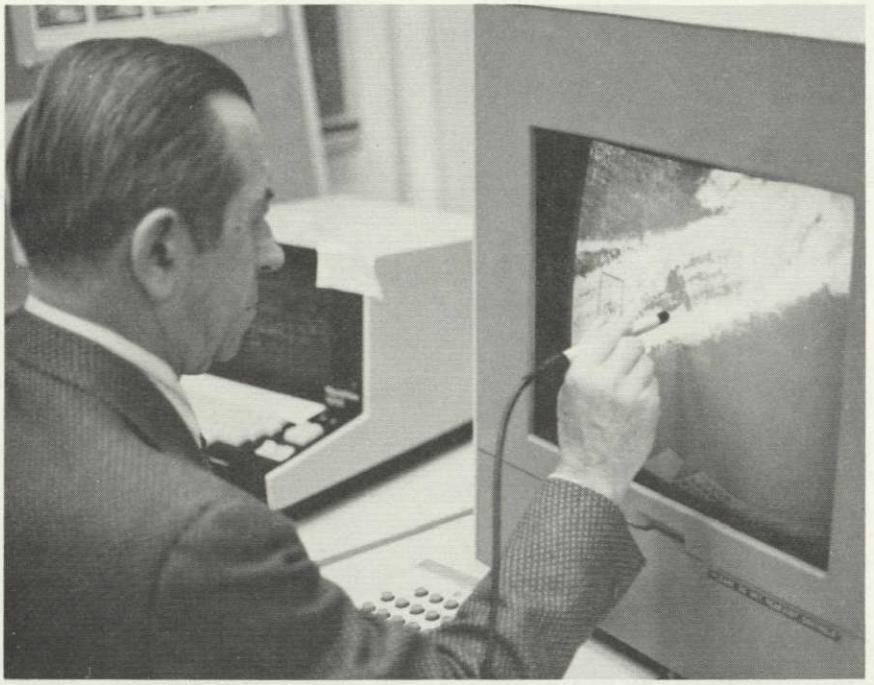


Figure 9 Analyst using a light pen to select training areas on a digital display

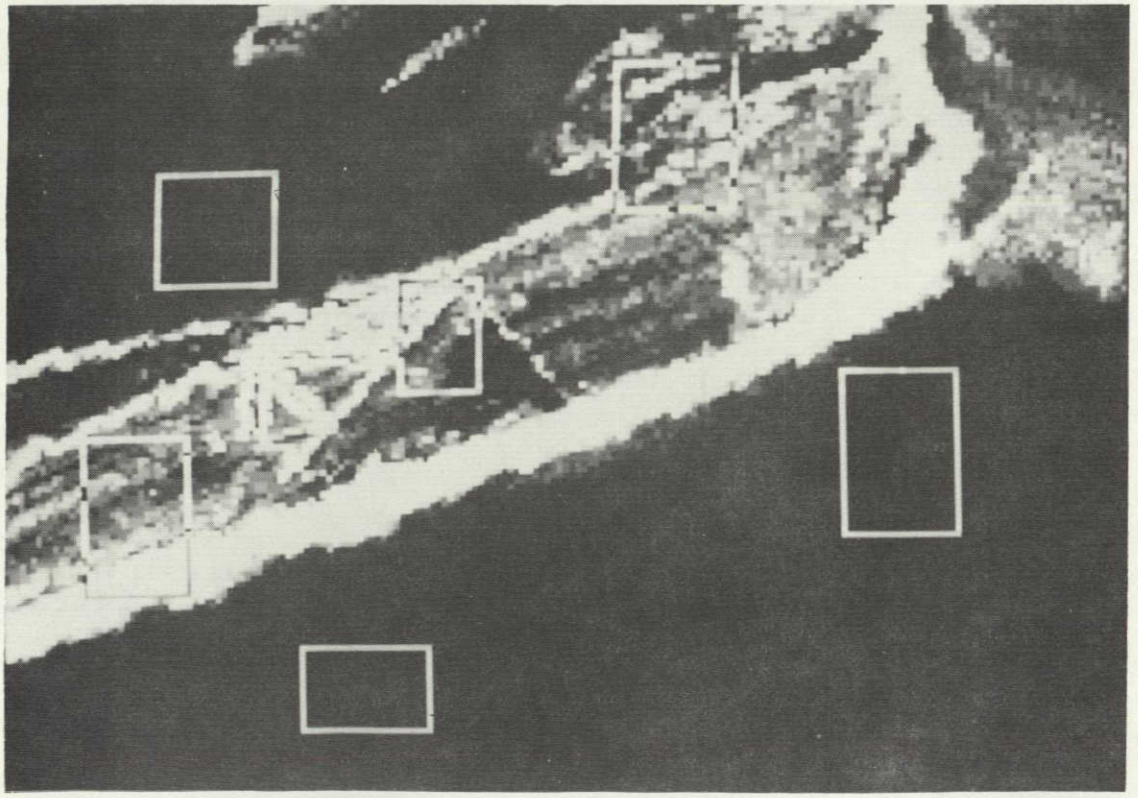


Figure 10 The six potential training areas selected from the study area

RAPI
STEVE

LABORATORY FOR APPLICATIONS OF REMOTE SENSING
PURDUE UNIVERSITY

PASS CAVALLO-WATER DATA, MAY 1973

RUN NUMBER..... 72072110
FLIGHT LINE... 112716260 TEXAS
DATA TAPE/FILE NUMBER.. 2781/ 1
REFORMATTING DATE. DEC 19,1975

DATE DATA TAKEN... NOV 27,1972
TIME DATA TAKEN..... 0926 HOURS
PLATFORM ALTITUDE..3062000 FEET
GROUND HEADING..... 180 DEGREES

CHANNEL 6 SPECTRAL BAND 0.60 TO 0.70 MICROMETERS CALIBRATION CODE= 1 CO = 0.0

THE CHARACTER SET USED FOR DISPLAY IS

BELOW	13.5		DISPLAYED AS W
FROM	13.5	TO 17.5	DISPLAYED AS W
FROM	17.5	TO 20.5	DISPLAYED AS $
FROM	20.5	TO 21.5	DISPLAYED AS X
FROM	21.5	TO 23.5	DISPLAYED AS /
FROM	23.5	TO 25.5	DISPLAYED AS *
FROM	25.5	TO 26.5	DISPLAYED AS I
FROM	26.5	TO 28.5	DISPLAYED AS /
FROM	28.5	TO 32.5	DISPLAYED AS =
FROM	32.5	TO 41.5	DISPLAYED AS -
FROM	41.5	TO 91.5	DISPLAYED AS
ABOVE		91.5	DISPLAYED AS

TRAIN FIELD BOUNDARIES MARKED WITH .
TEST FIELD BOUNDARIES MARKED WITH 1
SHARED FIELD BOUNDARIES MARKED WITH 1

HISTOGRAM BLOCK(S)

RUN NUMBER..... 72072110
LINES..... 1 45 154 11
COLUMNS.... 1 95 159 11
CALIBRATION CODE..... 1

LABORATORY FOR APPLICATIONS OF REMOTE SENSING
PURDUE UNIVERSITY

CAVALLO-WATER DATA, MAY 1973

RUN NUMBER..... 72072110
FLIGHT LINE... 112716260 TEXAS
DATA TAPE/FILE NUMBER.. 2781/ 1
REFORMATTING DATE. DEC 19,1975

DATE DATA TAKEN... NOV 27,1972
TIME DATA TAKEN..... 0926 HOURS
PLATFORM ALTITUDE..3062000 FEET
GROUND HEADING..... 180 DEGREES

CHANNEL 6 SPECTRAL BAND 0.60 TO 0.70 MICROMETERS CALIBRATION CODE= 1 CO = 0.0

THE CHARACTER SET USED FOR DISPLAY IS

BELOW	13.5		DISPLAYED AS W
FROM	13.5	TO 17.5	DISPLAYED AS W
FROM	17.5	TO 20.5	DISPLAYED AS $
FROM	20.5	TO 21.5	DISPLAYED AS X
FROM	21.5	TO 23.5	DISPLAYED AS /
FROM	23.5	TO 25.5	DISPLAYED AS *
FROM	25.5	TO 26.5	DISPLAYED AS I
FROM	26.5	TO 28.5	DISPLAYED AS /
FROM	28.5	TO 32.5	DISPLAYED AS =
FROM	32.5	TO 41.5	DISPLAYED AS -
FROM	41.5	TO 91.5	DISPLAYED AS
ABOVE		91.5	DISPLAYED AS

TRAIN FIELD BOUNDARIES MARKED WITH .
TEST FIELD BOUNDARIES MARKED WITH 1
SHARED FIELD BOUNDARIES MARKED WITH 1

HISTOGRAM BLOCK(S)

RUN NUMBER..... 72072110
LINES..... 1 45 154 11
COLUMNS.... 1 95 159 11
CALIBRATION CODE..... 1

MAY 15,1976
07:59:34 AM
LARSYS VERSION 3

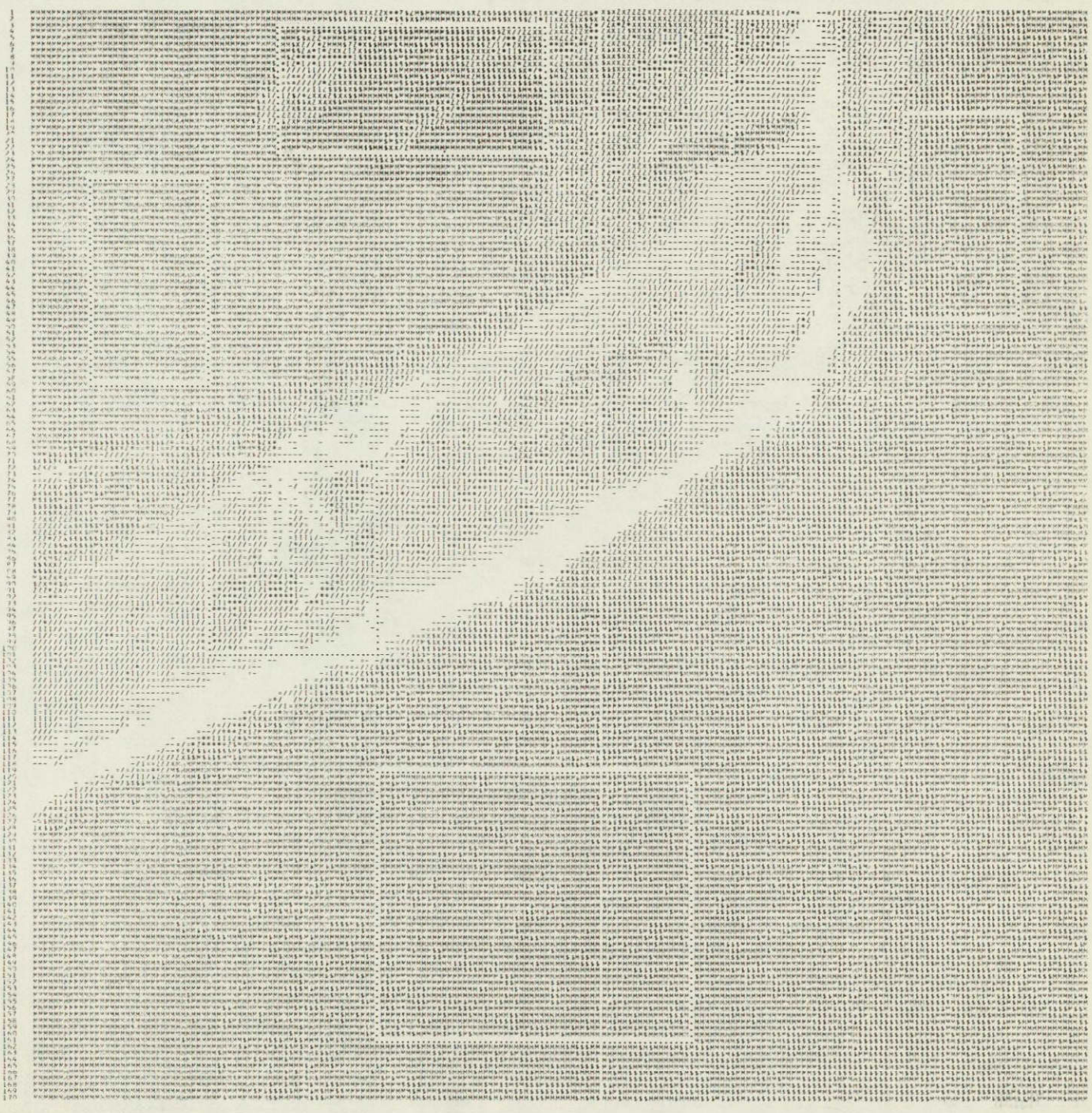


Figure 11 Gray scale printout of the quadrangle and the six training sites selected

arise at this time. The first question is, "Where in the area to be analyzed (in this case, the Pass Cavallo quadrangle) should the training areas be selected?" One obvious yet important consideration is the location of areas for which some reference data is available. In addition, the training areas should represent all of the variability present in the total area. This generally means that the training areas should be distributed throughout the entire area being analyzed.

To begin, the analyst looked at one of the pieces of reference data available, the topographic map, to acquaint himself with the area (see Figure 7). To select the initial training areas the analyst viewed the data on a digital display (see Figures 9 and 10). In this case, the analyst chose three land areas and three water areas as his initial training areas. A gray scale printout was then requested to determine the exact coordinates of the three land and three water areas selected (see Figure 11). Slight modifications of the exact coordinates of the training areas are often made at this point to include or exclude certain ground features by referring to the reference data (see Figure 12).

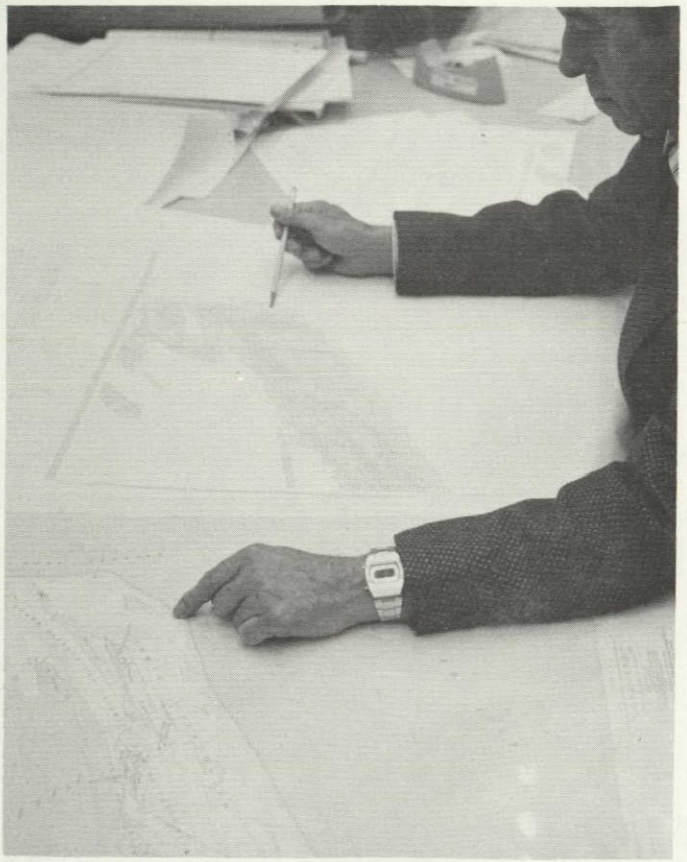


Figure 12 The analyst using reference data to check the location of training areas on the computer generated gray scale

In an effort to obtain representative training statistics, the areas selected include different ground features and various aquatic habitats and water depths.

Another question the analyst must ask himself is "What size training areas should be used?" Or, expressed another way, "How many data points are necessary to train the computer?" An analyst must consider theoretical lower limits and practical upper limits in answering this question. The lower limits are related to the number of features (wavelength bands) to be used in classifying the data. The minimum number of data points per training class necessary to carry out the mathematical calculations is n + 1 where n is the number of features. However, to get a good estimate of the required statistics, a more adequate number of data points per training class is on the order of 10n, where n is again the number of features. In general, a greater number of data points for training can be expected to provide more representative statistics, but there are practical upper limits to consider, too. The cost and availability of reference data and computer storage capabilities both place upper limits on the size of the training areas. In our example, the analyst was able to select 10,000 data points for each of the two general cover types (i.e. land and water). For example, he selected water areas of the following size:

	No. of Lines	No. of Columns
Water Area 1	20	30
Water Area 2	40	25
Water Area 3	40	45
	100	100

It should be pointed out that the number 100 has no special significance. By experience, the analyst has learned that a total of about 10,000 data points provide a good compromise between having enough data points to obtain good statistical representation of the classes of interest and few enough so that the computer time used in subsequent steps is not excessive.

The computer will classify every data point in the study area into one of the classes described by the training statistics. Therefore, the analyst should attempt to represent in the training data every cover type in the area, even ones having no direct informational value. For example, if you were investigating some characteristics of vegetation in a marsh region, you would not be interested in beach and sand dune areas located in the data set. However, if a training class for sand dunes is not provided, the computer will classify that area incorrectly because the CLASSIFYPOINTS algorithm requires that every data point be assigned to some class. Thus, even

class the classification will be in error because sand dunes were not included among the training classes.

A more thorough understanding of what training data is and why it is needed can be found in LARS Information Note 110474, An Introduction to Quantitative Remote Sensing by John Lindenlaub and James Russell.

Step 4-B Divide Training Areas into Subclasses

The next step in the development of training statistics is to determine if the training areas need to be divided into subclasses. The analyst accomplishes this by using the algorithm CLUSTER (see Appendix A, page 46). CLUSTER groups data vectors into clusters having similar spectral reflectance characteristics. When data are clustered, there is a tendency for the data points in each cluster or subclass to have a Gaussian or normal distribution. This use of clustering to find Gaussian subclasses is important because the classification algorithm to be used is based on a Gaussian assumption, i.e., that the distribution of the data to be classified can be adequately characterized by a set of Gaussian density functions.

For this analysis, three training areas from land and three from water were chosen, as described in the first part of this section. The analyst used the three land areas as input to one cluster analysis, and the three water areas in another.

Clustering is basically a computer operation with the analyst supplying only input information which tells the computer which multispectral image storage tape the data is on, the lines and columns defining the area to be clustered, and the number of clusters desired. The number of clusters requested by the analyst is influenced by the cover types of interest and their estimated spectral variability. The experience of the analyst plays an important role in his assessment of anticipated spectral variability. Once the cluster processor has found the desired number of clusters, it punches Field Description Cards which contain the line and column addresses of those data points included in each cluster.

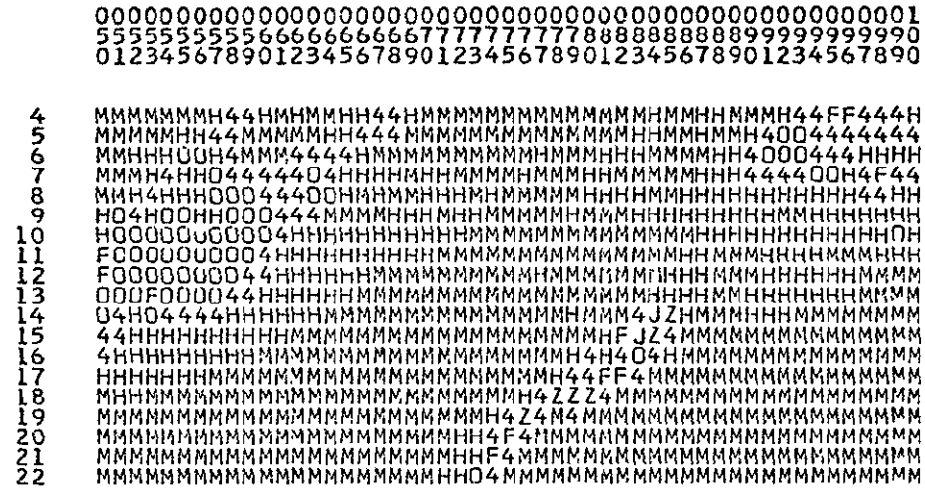
Thirteen clusters in the land areas and eight in the water areas were requested by the analyst for the analysis of the training areas for the Pass Cavallo quadrangle. The cover types of interest, as stated in the objectives, are various water classes, rangeland, marshes, brushland, barren land, and man-made features. By studying the aerial photography and from his knowledge of the area based upon field observations and other information, the analyst identified approximately 13 different types of ground cover and land features. An obvious feature from looking at the aerial photograph is the airport. The beach (both wet and dry) is another type of ground cover present. Reclaimed land (made from mud, sand and shell by man) and unvegetated mud flats were also evident. Within the land area there were also some ponds. The vegetated portion of the quadrangle was broken down into a number of ground cover types, including trees mixed with shrubs (low density trees), ridges covered with grass,

FIELD INFORMATION

FIELD LAND1
RUN NO. 72072110
OTHER INFORMATION

TYPE
NO. OF SAMPLES 969

LINES 4- 22 (BY 1)
COLUMNS 50- 100 (BY 1)



NUMBER OF POINTS PER CLUSTER

CLUSTER	1	2	3	4	5	6	7	8	9	10
SYMBOL	+	-	=	/	I	J	Z	8	F	0
POINTS	0	0	0	0	0	2	6	0	11	61

CLUSTER	11	12	13
SYMBOL	4	H	M
POINTS	86	233	570

Figure 13 CLUSTER printout for a land area within Pass Cavallo quadrangle

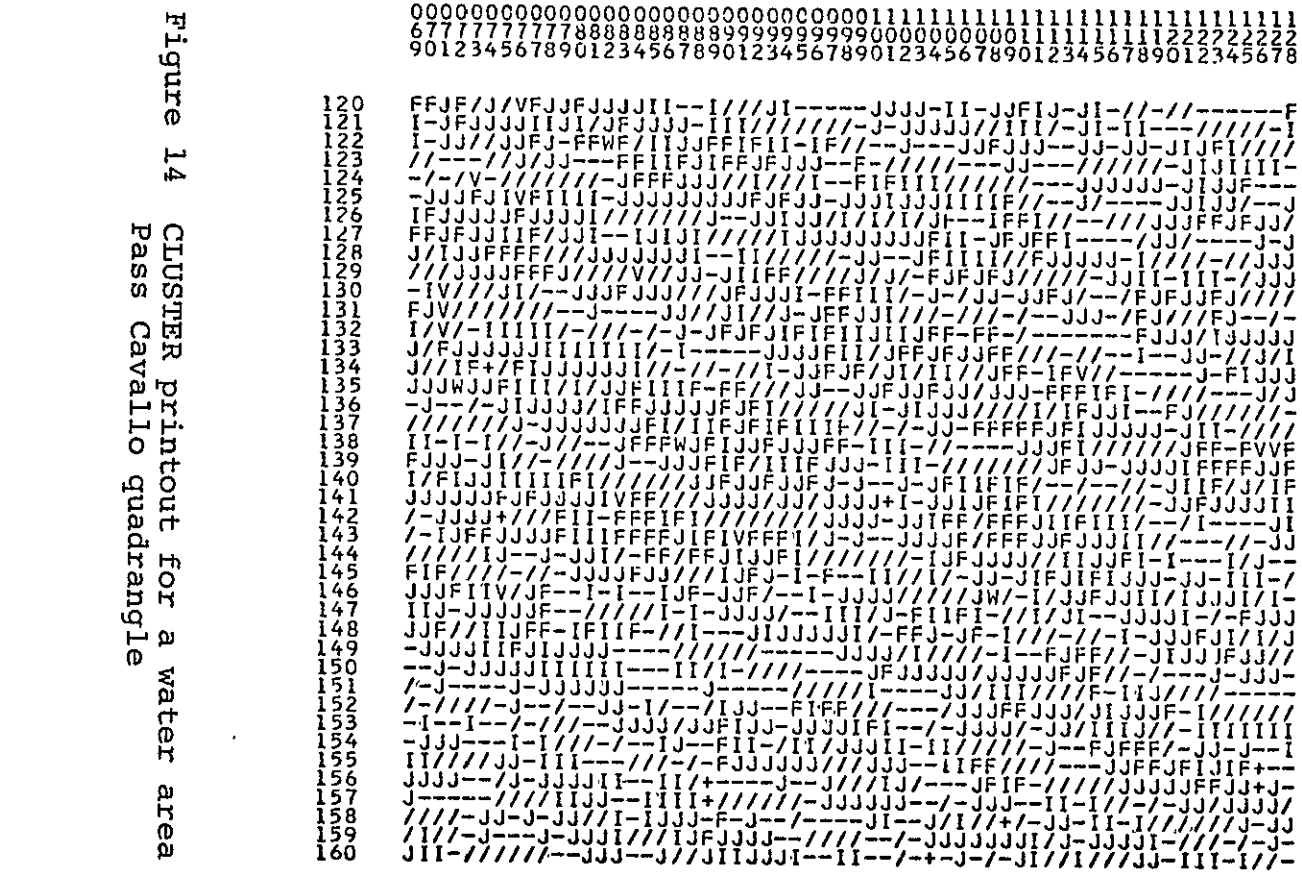
PASS CAVALLO-WATER DATA, MAY 1973

FIELD INFORMATION

FIELD WATER3
RUN NO. 72072110
OTHER INFORMATION

TYPE
NO. OF SAMPLES 2460

LINES 120- 160 (BY 1)
COLUMNS 69- 128 (BY 1)



NUMBER OF POINTS PER CLUSTER

CLUSTER	1	2	3	4	5	6	7	8
SYMBOL	+	-	/	I	J	F	V	W
POINTS	9	496	563	381	729	265	13	4

Figure 14 CLUSTER printout for a water area within Pass Cavallo quadrangle

vegetated flatlands and pasture. Another type of ground cover present was marshes, both salt water and fresh water. A CLUSTER printout for a portion of the land area of the Pass Cavallo quadrangle is shown in Figure 13.

The water covered portion of the quadrangle contains a number of spectrally distinct water classes. The difference in spectral reflectance from these areas may be due to such factors as depth, turbidity and surface motion or roughness. Based on his experience, the analyst chose to ask for 8 water cluster classes. The computer output for one of the water training areas is shown in Figure 14.

Step 4-C Calculate Statistics

The Field Description Cards describing the 13 clusters from land and 8 from water were then used by the analyst as input for calculating statistics for cover types in the training areas. The processing function known as STATISTICS (see Appendix A, page 48) was used for this purpose. To run STATISTICS the analyst must supply Field Description Cards for each information subclass, in this case they have been provided by CLUSTER. The program then retrieves data for these areas and calculates the mean vector and covariance matrix for each class and stores this information in a disk file or punches it on a deck of cards for use in subsequent steps in the analysis.

Self-Check

IV-A. How are training areas selected?

IV-B. What is clustering?

IV-C. How is clustering used in the development of training statistics?

IV-D. How are the statistics used in numerical analysis?

SECTION V CLASSIFY THE AREA

Upon completion of this section, you should be able to:

1. Describe the general procedure used by the computer in classifying an area.

Up to this point the effort of the analyst has been devoted to selecting and refining training data to be used by the computer for the classification process. The analyst has worked with reference data and has used the capabilities of the computer to produce products such as cluster maps to aid him in this task. The end product of his effort was the set of training statistics developed in the previous step of the analysis.

The classification step of the analysis is a "pure machine" operation. The analyst just has to supply to CLASSIFYPOINTS, the classification algorithm used in this step, the training statistics and coordinates of the area to be classified. The CLASSIFYPOINTS algorithm examines each data vector in the area of interest, assigns it to the class to which it most probably belongs and stores the result on a results file (see Appendix A, page 45). This process, known as a maximum likelihood classification, uses the training class mean vectors and covariance matrices (i.e., the training statistics) along with the data from each point to calculate the likelihood of that point belonging to each of the training classes. It then assigns the point to the most likely class. The algorithm also stores in the results file a second number for each data point. This number (the discriminant function value corresponding to the class to which the point was assigned) is a measure of the liklihood that the point actually belongs to this class -- a sort of confidence measure. These numbers are used in the next step of the analysis to determine which points should be thresholded. For further details on data classification, see Pattern Recognition: A Basis for Remote Sensing Data Analysis by Philip H. Swain (LARS Information Note 111572).

The analyst doesn't "see" the results of this step of the analysis because the end product is just a computer file containing class assignments and discriminant function values for each point in the area of interest. The analyst interacts with this file in the next step of the analysis.

Self-Check

V-A. In general how does the computer classify an area?

SECTION VI PRINT CLASSIFICATION MAP

Upon completion of this section, you should be able to:

1. Describe the utility of printing a classification map.

2. Define "thresholding."

Placing the classification results in a computer file allows the analyst to interact with the results in a variety of ways and permits him to generate a variety of output products. By assigning different symbols to each subclass, the analyst can have the computer produce a map showing the subclass to which each data vector was assigned. Percent of area in each class or subclass can be calculated directly (without producing a map) and the results printed in tabular form. Figures 15 and 16 show examples of the PRINTRESULTS processor for map and tabular output respectively. In each figure the 21 original spectral classes have been grouped into nine informational classes. The PRINTRESULTS processor (see Appendix A, page 47) also allows the analyst the flexibility of specifying that any data points within the quadrangle that did not have at least a predetermined minimum probability of belonging to the class to which they were assigned in the previous step be displayed as blanks. The process of not printing a symbol for the data points in question is called thresholding. Upon instructions from the analyst, the computer compares each data point's probability of belonging to its assigned class to a "threshold" value for the class. If this probability is less than the threshold value, the point is represented by a blank on the map. A threshold is specified on a class-by-class basis in terms of a percentage. The percentage indicates the approximate portion of points correctly classified into that class which the analyst is willing to have thresholded in an attempt to threshold all of the points incorrectly classified into that class. The strategy, then, is to pick threshold percentages (one for each class) which optimize the trade off between thresholding as few of the correctly classified points as possible but all or nearly all of the incorrectly classified ones. Thus, the percentage of data points actually thresholded can be larger or smaller than the requested value. Typically, analysts choose threshold values from 0% to 10%, and apply the same one to all classes.

RAP37
STEVE

LABORATORY FOR APPLICATIONS OF REMOTE SENSING
PURDUE UNIVERSITY

SEPT 14,1976
08 35 30 AM
LARSYS VERSION 3

CLASSIFICATION STUDY 625777306
RUN NUMBER..... 72072104
FLIGHT LINE... 112716260 TEXAS
DATA TAPE/FILE NUMBE.. 23767 1
REFORMATTING DATE. FEB 27,1975

CLASSIFIED SEPT 13,1976
DATE DATA TAKEN... NOV 27,1972
TIME DATA TAKEN..... 3926 HOURS
PLATFORM ALTITUDE..3062000 FEET
GROUND HEADING..... 180 DEGREES

CLASSIFICATION TAPE/FILE NUMBER ... 8/ 17

CHANNELS USED

CHANNEL 5 SPECTRAL BAND 0.53 TO 0.60 MICROMETERS CALIBRATION CODE = 1 CO = 0.0
CHANNEL 6 SPECTRAL BAND 0.60 TO 0.70 MICROMETERS CALIBRATION CODE = 1 CO = 0.0
CHANNEL 7 SPECTRAL BAND 0.70 TO 0.80 MICROMETERS CALIBRATION CODE = 1 CO = 0.0
CHANNEL 8 SPECTRAL BAND 0.80 TO 1.10 MICROMETERS CALIBRATION CODE = 1 CO = 0.0

CLASSES

SYMBOL	CLASS	GROUP	SYMBOL	CLASS	GROUP
+	1	BEACH-DRY	1	12	VEG-SWALES
-	2	BEACH-WET	=	13	AIRPT/DUNES
=	3	AIRPT/DUNES	#	14	VEG-RIDGES
M	4	VEG-RIDGES	*	15	VEG-RIDGES
I	5	VEG-SWALES	C	16	WATER B
I	6	VEG-SWALES	/	17	WATER A
M	7	VEG-RIDGES	C	18	WATER B
I	8	VEG-SWALES	L	19	WATER C
C	9	VEG-SUBMERG	L	20	WATER C
I	10	VEG-SWALES	/	21	WATER A
*	11	VEG-RIDGES			

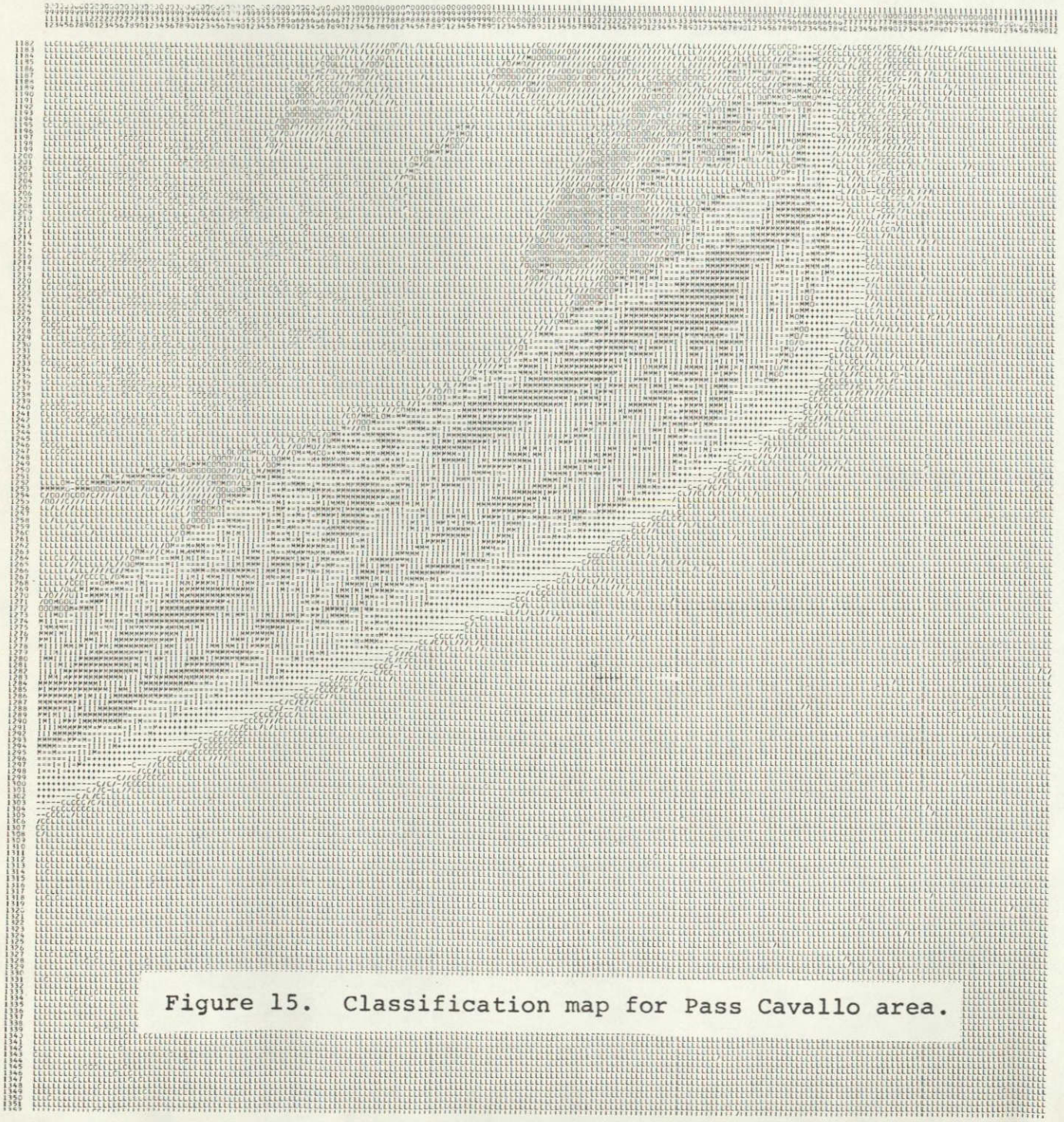


Figure 15. Classification map for Pass Cavallo area.

LABORATORY FOR APPLICATIONS OF REMOTE SENSING
PURDUE UNIVERSITY

CLASSIFICATION STUDY 625777306
RUN NUMBER..... 72072104
FLIGHT LINE... 112716260 TEXAS
DATA TAPE/FILE NUMBER.. 2376/ 1
REFORMATTING DATE. FEB 22,1975

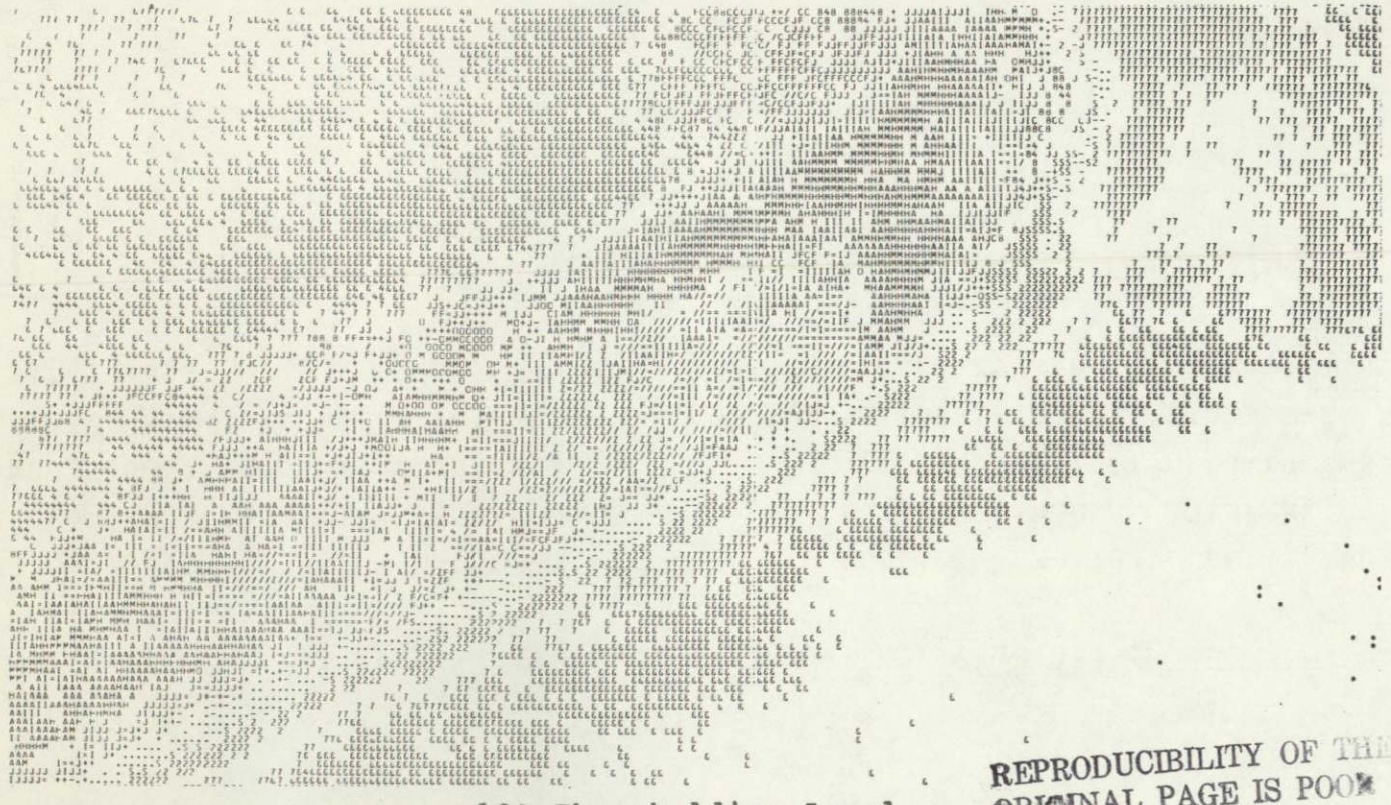
CLASSIFIED SEPT 13,1976
DATE DATA TAKEN... NOV 27,1972
TIME DATA TAKEN..... 0926 HOURS
PLATFORM ALTITUDE..3062000 FEET
GROUND HEADING..... 180 DEGREES

CLASSIFICATION TAPE/FILE NUMBER ... 8/ 17

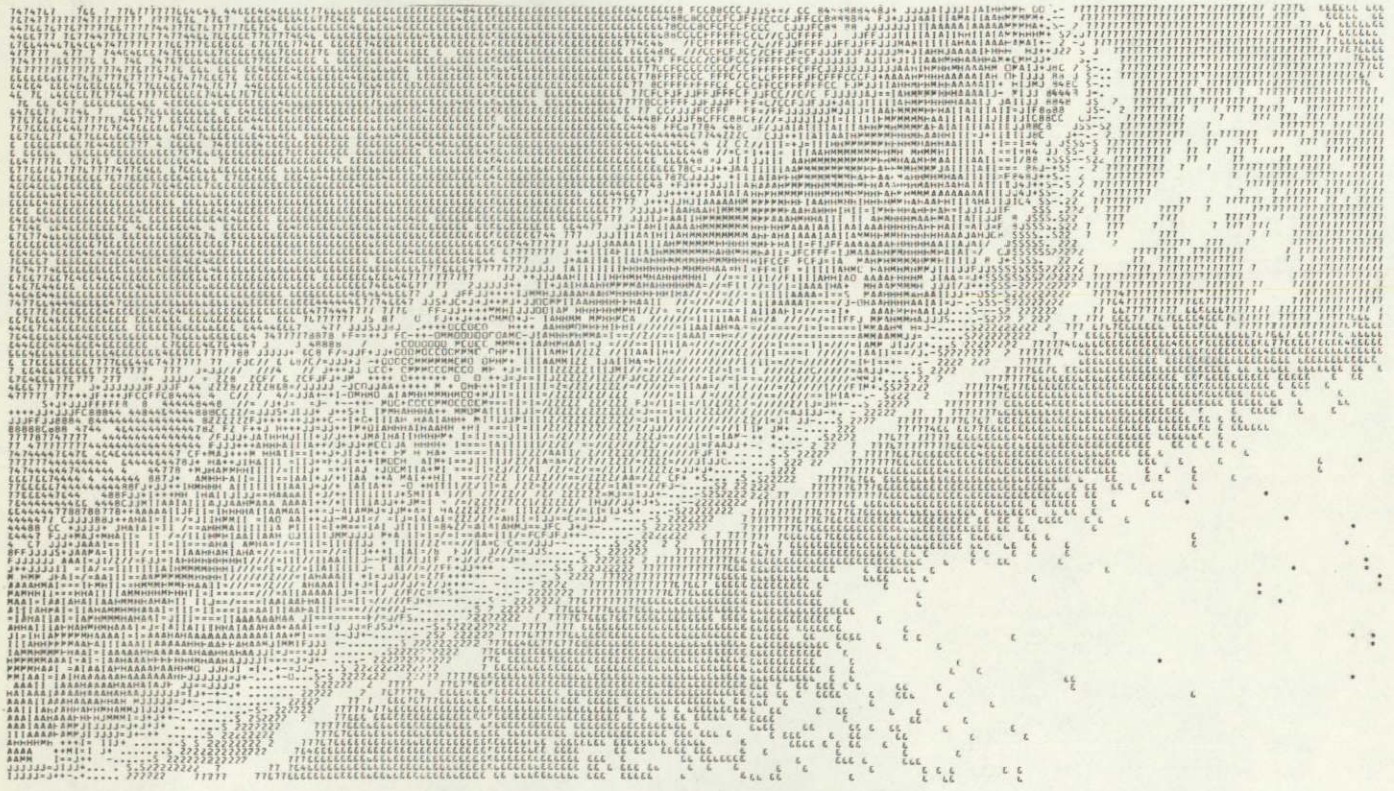
SYMBOL	GROUP	POINTS	ACRES	HECTARES	PERCENT
M	VEG-RIDGES	2395	2634.5	1066.6	6.5
I	VEG-SWALES	1943	2137.3	865.3	5.3
+	BEACH-DRY	746	820.6	332.2	2.0
-	BEACH-WET	434	477.4	193.3	1.2
O	VEG-SUBMERG	921	1013.1	410.2	2.5
=	AIRPT/DUNES	992	1091.2	441.8	2.7
/	WATER A	1276	1403.6	568.3	3.5
C	WATER B	1593	1752.3	709.4	4.3
L	WATER C	26464	29110.1	11785.5	72.0
	TOTAL	36764	40440.0	16372.5	100.0

EACH DATA POINT REPRESENTS 1.10 ACRES
0.45 HECTARES

Figure 16. Tabular output for Pass Cavallo area.



10% Thresholding Level



1% Thresholding Level

Figure 17. Comparison of two thresholding levels for a portion of Pass Cavallo quadrangle

The results of applying two different sets of threshold levels are shown in Figure 17. The classification map on the top was made using a threshold value of 10% while the map on the bottom was done with a threshold value of 1%. The analyst in this land use example has found from experience that a threshold value of 0.5% is an appropriate value to use. Applying threshold values to the classification map "flags" those points with a low probability of actually belonging to one of the initial training classes. The result is a map which shows as white or blank those areas of the data which are not adequately represented by any of the training classes. How this threshold map is used to refine the training statistics is discussed in the next section.

Self-Check

VI-A. Why is a classification map printed?

VI-B. What is "thresholding?"

SECTION VII EVALUATE CLASSIFICATION PERFORMANCE

Upon completion of this section, you should be able to:

1. Briefly discuss the process of refining the training data using thresholded areas.

2. Describe three methods for checking classification performance.

Evaluation of classification performance is an important step in the analysis sequence. Referring to the flow chart (Figure 8, page 18), you can see that this is a decision point in the process. The decision to be made is whether or not the classification of the cover types for the entire area is adequate. The first step in answering this question involves looking at the thresholded areas of the PRINTRESULTS map. If the size of the thresholded areas is large compared to the fields, forests, lakes, or residential areas being mapped, then it is likely that ground cover in those areas was not represented in the training statistics. If so, the analyst selects additional training areas from within the white (blank) portions of the classification map where the classifier results have been "thresholded."

The analyst uses reference data and knowledge of the reflectance characteristics of earth surface features to select training data and label them as to the ground cover they represent. The analyst again uses clustering to divide the new training areas into subclasses. When the training areas found by thresholding have been clustered and the Field Description Cards punched, the STATISTICS processing function is run again. This time the input data deck is the original set of Field Description Cards from CLUSTER plus the cards produced by clustering the data selected from the thresholded areas. The new set of statistics is used to re-classify the quadrangle.

Typically, this classification will again have thresholded areas from which more training samples may be selected. Again statistics may be calculated and the area classified. The analyst then evaluates the classification results and decides on the basis of training class performance and the amount of thresholding, whether to terminate or further refine the analysis. When training class performance is high, the training classes are spectrally separable; when very few points are thresholded, the training classes are representative of the entire scene. If the results are still unsatisfactory, he again selects training samples from the thresholded areas. If the results are satisfactory, he moves on to apply them. Different analysis problems require different levels of classification accuracy. Also, the

mechanisms for determining the accuracy will vary somewhat from one analysis to another.

One technique is to obtain training class performance. Recall that training classes are a collection of data points which have been identified by the analyst as being a certain type of ground cover. These points were used to "train" the computer earlier. Now these same points are checked to see if they were classified into the class the analyst placed them in. The number correctly classified is expressed as a percentage and is referred to as training class performance. Training class performance does not indicate whether the classes have been correctly assigned to the specific cover types, but only whether the classes are spectrally distinct. For example, the classes may be spectrally distinct, but not represent the ground covers of interest. Reference data, such as B.E.G. maps, can be used to correlate the spectral classes with actual ground cover types.

Another method of evaluating performance involves selecting coordinates for homogeneous areas which are called "test fields." These are selected by the analyst to represent known cover types in the analysis area. Some type of reference data (B.E.G. maps, topographic maps, ground observations, or photography) is used to choose the test fields. The computer is then requested to tabulate the number of points accurately classified in the test fields. The percent correct is used as one criterion for deciding if there is a need to return to the analysis steps in Section IV.

A third technique for evaluating performance is to use statistically random or stratified procedures to select areas over which to make actual ground observations to check the classification results. This procedure can be the most valuable method of checking accuracy of the classification, providing observations are made soon after the analysis. Photographs are also useful at this point but subject to human interpretation and error.

The technique you select will be dependent on the availability of reference data. A combination of these techniques will usually produce the most satisfactory evaluation of the classification results.

If the results of the evaluation are satisfactory, the analyst moves to the application of the results (described in Section VIII). If the results are not satisfactory, the analyst needs to refine his training statistics as described in Section IV.

Self-Check

VII-A. How are thresholded areas used to refine training data?

VII-B. Describe two of the three methods for checking classification performance.

SECTION VIII APPLY RESULTS

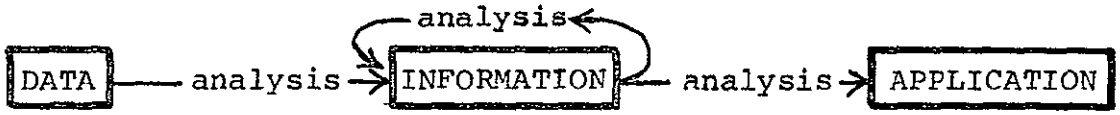
Upon completion of this section, you should be able to:

1. Name at least two types of information that can be extracted from a land use classification.

2. Name at least two types of applications that can be made from the results of a land use classification.

3. Describe an example of useful information extracted from the analysis of remotely sensed data in your particular field of land use.

An overall perspective of the sequence described in this simulation is shown here:



So far we have concentrated on obtaining data and analyzing it to obtain information. We are now ready to analyze the information in terms of the proposed application(s). Several formats of information have been obtained as a result of analyzing the data. We have met our initial objectives by producing a general land use map and determining the extent of specific cover types. The final and perhaps most important step is the interpretation and application of this information.

Two possible applications for the information produced in our example analysis include:

1. Update the Environmental Geology Sheets (1:125,000 scale) contained in the University of Texas, Bureau of Economic Geology Atlas of the Texas Coastal Zone.

2. Monitor and evaluate wildlife habitats within the area covered by the Pass Cavallo 7½ minute USGS quadrangle.

This is certainly not an exhaustive list of applications. Each area is different and the analysis objectives may vary widely. The advantage of the computer-aided analysis can be seen when one considers the time savings in man hours spent surveying a large area. Repetitive surveys are also possible since LANDSAT data are collected over the same area every eighteen days.

However, let us consider the objective stated in the second hypothetical application. The classification map (Figure 15, Section VI) of the Pass Cavallo quadrangle shows that the area is dominated by Matagorda Island and the surrounding coastal waters. Tabular statistics (Figure 16, Section VI) indicate that the quadrangle area is composed of 80% water, 18% land and 2% shallow water with submerged vegetation.

In considering wildlife habitats, we will probably only be interested in the 20% land and submerged vegetation areas. An analysis of the area may show that Matagorda Island is being used for leased grazing as well as wildlife habitat. Again looking at Figure 16 we find that Matagorda Island has the following proportion of classes:

Class	Acres	
Beach Area	1298	
Wet		477
Dry		821
Vegetation	5785	
Ridges		2635
Swales		2137
Submerged		1013
Airport and Dunes	1091	

Comparison of these figures to information collected earlier may show that leased grazing is leading to a decrease in vegetative cover due to over grazing. This loss of vegetated areas could decrease wildlife nesting areas, thus reducing the amount of productive habitat. One possible use of this information would be to limit, or prohibit entirely, leased grazing on Matagorda Island to allow for the protection and regeneration of wildlife habitat.

Examples of results from multispectral data classification with application to land use may be found in several journals, including:

Journal of Soil and Water Conservation
Photogrammetric Engineering and Remote Sensing
Remote Sensing of the Environment
Soil Science Society of America Proceedings
Agronomy Journal

Self-Check

VIII-A. What types of information can be extracted from a land use classification?

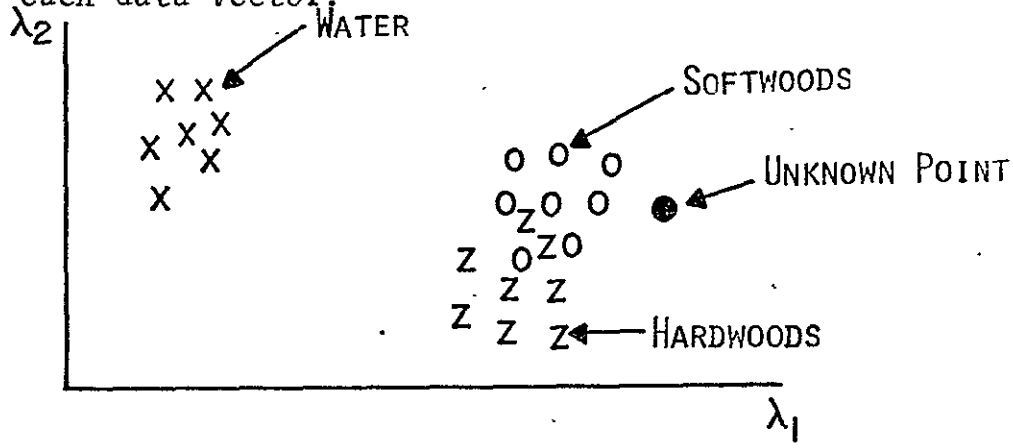
VII-B. What applications can be made from the results of a land use classification?

VIII-C. Describe an example in your field of the type of information which can be extracted from the analysis of remotely sensed data.

APPENDIX A

SUMMARY OF RELEVANT LARSYS PROCESSING FUNCTIONS

CLASSIFYPOINTS performs the maximum likelihood classification on a point-by-point basis over an area specified by the user. As viewed in the multidimensional data space CLASSIFY-POINTS establishes the decision boundaries used to classify each data vector.



Input

- Data from Multispectral Image Storage Tape
- Control Cards to select the processing and output options
- Field Description Cards indicating area(s) to be classified.
- Statistics for each training class

Process

The program uses the training class mean vectors and covariance matrices and the data from each point to calculate the probability that the point belongs to each of the training classes. It then assigns the point to the most probable class.

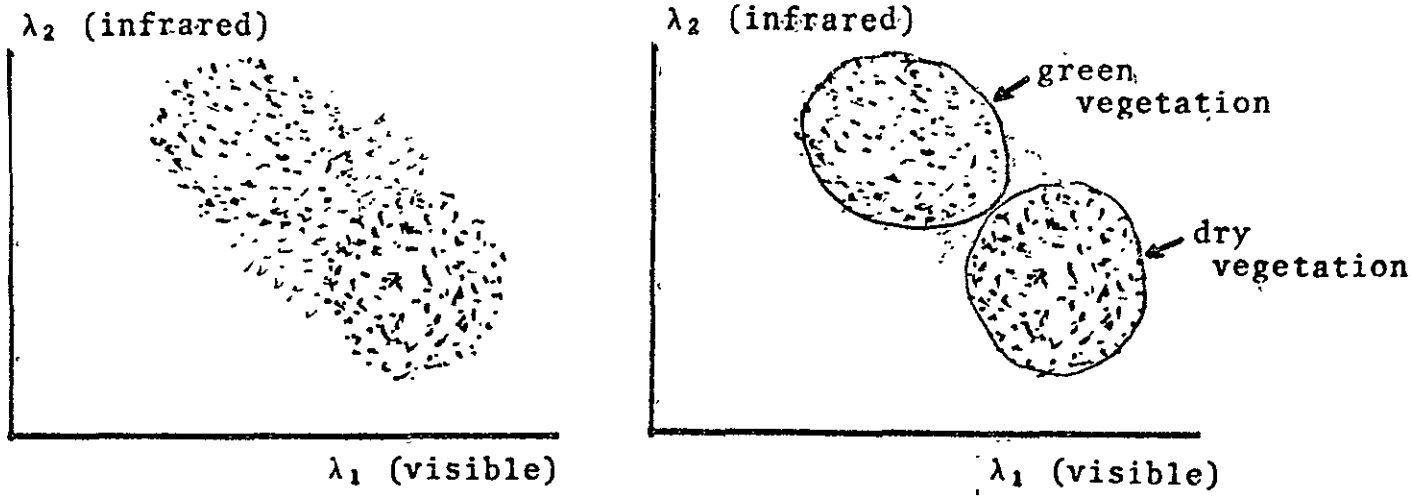
Output

- A Classification Results File normally used as input to the PRINTRESULTS processing function which produces a variety of printed output

44

PRECEDING PAGE BLANK NOT FILMED

CLUSTER is a process that groups individual data points into a predefined number of groups (clusters) specified by the analyst.



Input

- Data from Multispectral Image Storage Tape
- Indication of the Area to be Clustered
- Number of Clusters Desired

Process

1. The computer assigns a location in the feature space as the initial center of each cluster.

2. It then calculates the distance between each data point and each cluster center and assigns the point to the cluster with the minimum distance.

3. Next, new cluster centers are determined by calculating the mean vector for the data points assigned to each cluster.

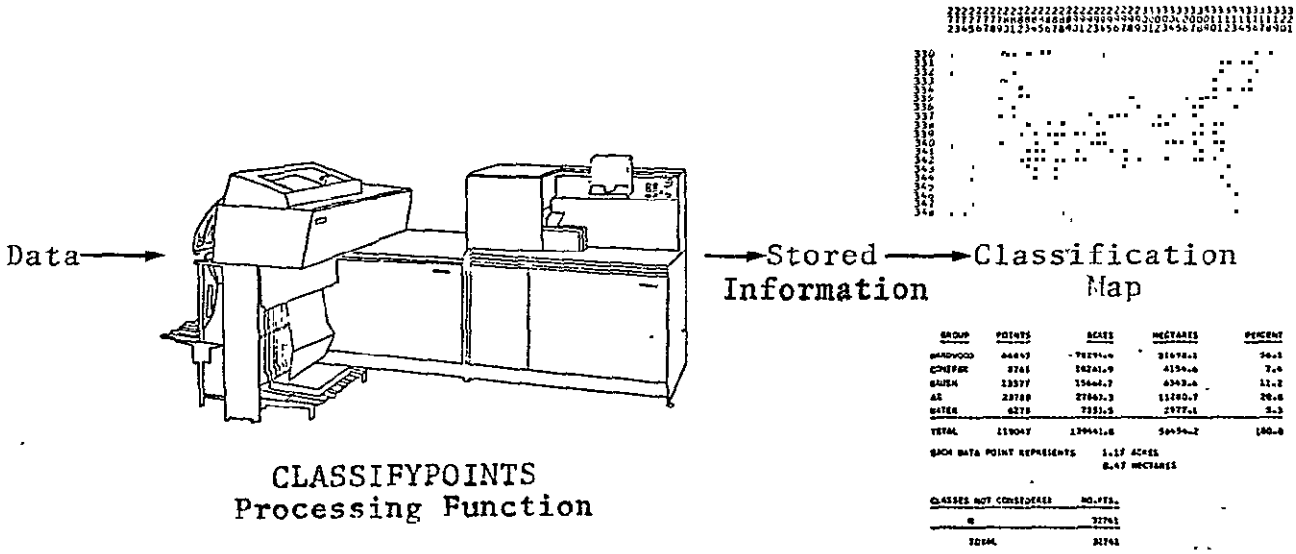
4. The computer then proceeds back to Step 2 and reassigns each sample to the closest newly defined cluster center.

5. The computer continues the cycle of calculating the cluster centers (Step 3) and reassigning data points (Step 2) until all data points (or an analyst specified percentage of the data points) are assigned to the same cluster in a succeeding cycle.

Output

- A Cluster Map which pictorially represents each area that was clustered
- A summary of the number of points assigned to each cluster
- Tabular output, statistics and Field Description Cards

PRINTRESULTS produces a variety of printed outputs describing the results of a classification in the form of a map and/or tabular output.



Input

- Location (in computer file) of classification results to be printed.
- Symbols to be assigned to various classes.
- Area(s) to be used for maps and tables.

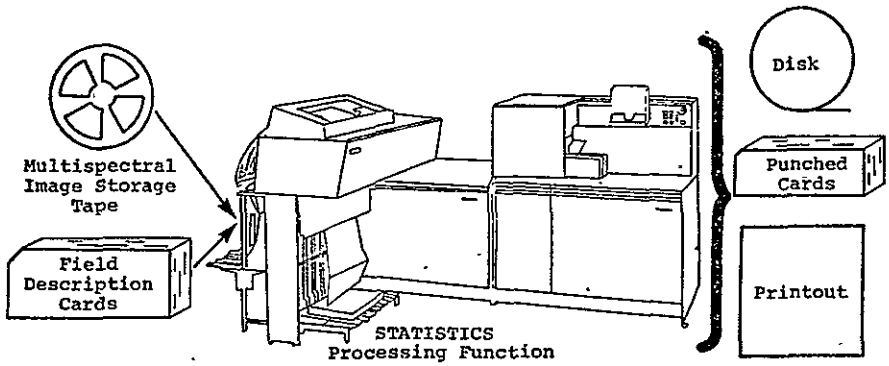
Process

The process prints out the information stored in the computer that was produced by the CLASSIFYPOINTS processing function.

Output

- Classification map (using specified symbols) with outline of training and test fields (if requested).
- Tables showing training field and class performance.
- Tables showing test field and class performance.

STATISTICS is a processing function that calculates the mean vectors and covariance matrices for training fields and/or training classes based on the multispectral data in channels specified by the user.



Input

- Data from Multispectral Image Storage Tape(s)
- Field Description Cards to define Training Fields

Process

Each **Field Description Card** specifies a single training field by line and column coordinates. The channels for which statistics are to be computed must always be specified.

From the data values stored on the Multispectral Image Storage Tape the computer calculates the statistics for the areas identified by the field description cards.

The results of the STATISTICS processing function are stored in the computer and may be punched on cards. The results are used by several other processing functions in the course of an analysis.

Output

- Statistics summary containing the mean and standard deviation vectors and correlation matrices of the channels for each field and/or class.
- Histograms of the data values for requested channels for each field and/or class.
- Spectral plots which show the normalized range of values for the requested channels for each field and/or class.

- Coincident spectral plots for specified groups of classes.

- Statistics deck which contains mean vectors, covariance matrices, Field Description Cards and channels used.

APPENDIX B

Answers to Self-Check Items:

I-A. In any order:
1) Location of area 2) Cover types of interest
3) Application of results 4) Classification performance expected

I-B. Your objective should be useable by you and contain the four basic parts listed above.

II-A. Aircraft and Satellites

II-B. Any three of the following:
Obstruction of the ground scene by 1) haze, 2) clouds or 3) snow cover. 4) Striping when one or more of the data lines is not properly recorded due to problems with the detectors and associated electronics.

II-C. 1) Rotating image
2) Changing scale

III-A. Any three of the following:
1) Maps 2) Aerial Photographs 3) Previous Analysis Results 4) Observations in the Field

III-B. The analyst compares remotely-sensed data with reference data in order to gain familiarity with the area, to identify ground cover types, to assist in developing training statistics and to help evaluate the classification results.

IV-A. Training areas are selected based upon the area to be analyzed. They are usually selected where there is reference data available. In addition, they must be representative of the total area (usually scattered throughout the area).

The minimum number of data points required for training is determined by the number of wavelength bands used for classification of the data (at least ten times the number of wavelength bands), the availability of reference data and computer storage capabilities. Of course, cost enters into the latter two considerations.

50

IV-B. Clustering is the computer algorithm which combines data values into groups with similar spectral characteristics.

IV-C. Clustering is used to combine the data points into groups based upon their spectral characteristics. The number of groups is determined by the cover types of interest--the more cover types present, the greater the number of clusters requested. Data in these clusters are then used to calculate the statistics for the various cover types.

IV-D. The statistics are used to train the computer to classify unknown points within the area being analyzed.

V-A. For each data vector in the area being classified the computer computes the likelihood of the data vector belonging to each of the training classes. It then assigns the data vector to the most likely class. Both the class assignment and the likelihood value are recorded in a computer file for use in the next step of the analysis process.

VI-A. A classification map is printed to provide a visual image of the classification results.

VI-B. Thresholding is the process of identifying those data points which do not have a minimum probability of belonging to the class to which they were assigned.

VII-A. Thresholded areas are used to determine that proportion of the total area that does not have a minimum probability of being classified accurately. If the size of the thresholded areas is large compared to the size of the ground features that you are trying to map, additional training areas are selected within the thresholded area in order to refine the train data.

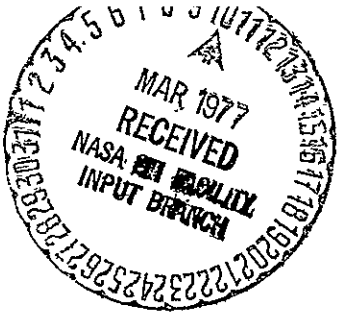
VII-B. Any two of the following:
1) Test fields are identified which represent known cover types and then the accuracy with which the computer classified these "known" areas is determined.
2) Training class performance uses the same collection of data points which was used to train the computer. The classification is then evaluated on the basis of how well the computer classified these training areas.

3) Random or stratified areas may be selected from the classification data and then checked by actual on-the-ground observations.

VIII-A. Information extracted from land use classification:
1) Maps showing various cover types and their location
2) Tabular charts indicating percentage of area in various cover types.

VIII-B. Applications of land use classification:
1) Land use mapping
2) Land use management
3) Land use planning
4) Monitoring change in areas

VIII-C. Many answers are appropriate depending upon your specific interests. The best answer is one that proves useful to you! (Many examples are included in the journals listed on page 42).


MAR 1977
RECEIVED
NASA
INPUT BRANCH